A study of lip movements during spontaneous dialog and its application to voice activity detection

David Sodoyer, Bertrand Rivet, Laurent Girin, Christophe Savariaux, Jean-Luc Schwartz
GIPSA–lab, department of Speech and Cognition, UMR 5126 CNRS, Grenoble-INP, Université Stendhal, Université Joseph Fourier, 46, av. Félix Viallet 38031 Grenoble, France.

Christian Jutten
GIPSA–lab, department of Images and Signal, UMR 5126 CNRS, Grenoble-INP, Université Stendhal, Université Joseph Fourier, 46, av. Félix Viallet 38031 Grenoble, France.

Third submission after revision – Date: 23 September 2008

Running title: Voice activity detection based on lip movements

This paper presents a quantitative and comprehensive study of the lip movements of a given speaker in different speech / non speech contexts, with a particular focus on silences (*i.e.*, when no sound is produced by the speaker). The aim is to characterize the relationship between “lip activity” and “speech activity”, and then to use visual speech information as a Voice Activity Detector (VAD). To this aim, an original audio-visual corpus was recorded with two speakers involved in a face-to-face spontaneous dialog, although being in separate rooms. Each speaker communicated with the other using a microphone, a camera, a screen, and headphones. This system was used to capture separate audio stimuli for each speaker and to monitor each speaker’s lip movements in synchrony with the recorded sound. A comprehensive analysis was carried out on the lip shapes and lip movements corresponding to either silence sections or non-silence sections (*i.e*. speech + non-speech audible events). A single visual parameter, defined to characterize the lip movements, was shown to be efficient for the detection of silence sections. This results in a Visual VAD (V-VAD) that can be used in any kind of environment noise, including intricate and highly non-stationary noises, *e.g.*, multiple and/or moving noise sources or competing speech signals.

PACS number: 43.72.-p Speech processing and communication systems

I. INTRODUCTION

A. Context: audio-visual speech processing

Speech is a bimodal signal, both acoustic and visual. Many studies have shown that the visual modality improves the intelligibility of speech in noise when switching from the “audio only” condition to the “audio + speaker’s face” condition (Sumby and Pollack, 1954; Erber, 1975; Benoît *et al*., 1994; Robert-Ribes *et al.*, 1998). In parallel, McGurk and McDonald (1976) demonstrated that humans can even integrate conflicting audio and visual speech stimuli to perceive a “chimeric” speech stimulus. More recently, Grant and Seitz (2000) have shown that viewing the speaker’s face also improves the detection of speech in noise. Such results have been confirmed by Kim and Davis (2004) and Bernstein *et al.* (2004). More specifically, visual information helps pertinent acoustic features to be better extracted, *i.e*., “seeing to hear better”, providing a different and complementary contribution to lip-reading (Schwartz *et al.*, 2004). Additionally, visual speech information has been shown to irresistibly attract speaker’s localization (Bertelson, 1999).

Concerning the nature of visual speech information, two major questions have been addressed. Firstly, the oral region including the lips and jaw seems to be the major contributor to visual speech perception (see, e.g., Summerfield, 1979; Benoît *et al.*, 1996). Thomas & Jordan (2004) actually showed that the intelligibility of oral-movements display was more or less the same as that of whole-face movements display. However, extra-oral movements also influence identification of visual and audiovisual speech, mostly due to the strong correlation between oral and extra-oral movements (Munhall and Vatikiotis-Bateson, 1998). Orofacial configurations can be basically characterized in terms of lip contours and specifically by the parameters of inner lip height, inner lip width and lip protrusion (Summerfield, 1979; Abry and Boë, 1986; Benoît *et al.*, 1992, 1996). Secondly, the question of static vs. dynamic processing of facial configurations has been largely discussed. Studies using point-like displays, which remove fine spatial information, showed that movement seems to be crucial in the perceptual processing of visual speech in both noisy configurations (Rosenblum et al., 1996) and conflicting McGurk stimuli (Rosenblum and Saldana, 1996). This led Munhall et al. (1996) to suggest that listeners might use the time-varying properties of visual speech for perceptual grouping and phonetic perception. Neurophysiological data seem to confirm the specific role of the dynamic processing of visual speech (Calvert & Campbell, 2003; Munhall et al., 2002). This is compatible with Summerfield (1987)’s suggestion that one possible metric for audiovisual integration is the pattern of changes over time in articulation, considering that listeners are sensitive to the dynamics of vocal tract change. Thereafter, a number of studies in the audio-visual speech literature have characterized the correlation between lower face movement and the produced acoustic signal (Yehia *et al.*, 1998; Barker and Berthommier, 1999; Jiang *et al.*, 2002; Bailly and Badin, 2002; Goecke and Millar, 2003).

Following these considerations on the bimodal aspect of speech, an important number of technological studies have been undertaken in the last twenty years to integrate the visual modality into speech processing systems. The goal is to improve the performance and robustness (in noise) of different human-to-human telecommunication systems or human-computer interfaces (HCI). Petajan (1984) was the first to integrate visual speech information in an automatic speech recognition (ASR) system. Many studies followed, including recent advances going towards real-life implementations of bimodal ASR (Potamianos *et al.*, 2003). Recently, audio-visual speech processing applications also concerned video indexing and retrieval (Huang *et al.*, 1999; Iyengar and Neti, 2001), audiovisual speech synthesis and talking heads (Yehia *et al.*, 2000; Bailly *et al.*, 2003; Cosi *et al.*, 2003; Gibert *et al.*, 2005), and audio-visual speech coding (Rao and Chen, 1996; Girin, 2004). In recent years, the visual modality has also been exploited for speech enhancement in (background) noise (Girin *et al.*, 2001; Deligne *et al.*, 2002; Potamianos *et al.*, 2003b), and more generally for speech source separation, *i.e.*, for the extraction of a speech signal from complex mixtures using several microphones, for both linear instantaneous mixtures (Sodoyer *et al.*, 2002, 2004) and convolutive mixtures (Wang *et al.*, 2005; Rivet *et al.*, 2007).

B. Video characterization of silence vs. non-silence sections

Most of the time, studies addressing the characterization of lip patterns in speech production have been carried out in more or less controlled speech production contexts (typically “laboratory speech”: see, e.g., Abry and Boë, 1986; Benoît *et al.*, 1992; Goecke and Millar, 2003; Jiang *et al.*, 2002; Yehia and Vatikiotis-Bateson, 1998). Relatively poor attention has been paid to the description and characterization of these patterns during speech production in natural contexts, especially in spontaneous multi-speaker conversation. Moreover, in such context, speech activity (*i.e.* actual speech production by a speaker of interest) alternates with many silence sections (*i.e.* sections where the speaker of interest does not produce sounds, whereas other speakers may actually do), and also with many non-speech audible events such as murmurs, grunts, laughs, respiration intakes, expirations, lip noise, whispers, sighs, growls, moans, *etc* (Campbell, 2007). In spite of this, even poorer attention has been paid to lip patterns in silence and non-speech contexts, although these patterns may exhibit a specific behavior, to be considered in both audio-visual speech fundamental studies and technological applications.

This paper provides an attempt to fill this gap. The relationship between a speaker’s lip movements and speech activity (or non-speech oral production) *vs.* silence is investigated, using signals from a spontaneous dialog. For this aim, the recording and the study of a “real-life” audio-visual corpus were achieved and are presented in this paper. This corpus consists of two speakers recorded in a spontaneous dialog situation (in French) during about 40 minutes. It is characterized by two properties. Firstly, it is based on a very clean audio (and of course video) recording process, since each speaker is located in a separate room to completely avoid cross-speaker audio interferences in the recordings. Communication between the two speakers is effected using a specially-designed equipment described in Section II.A. Secondly, the audiovisual material is recorded in a lively dialog situation, in which various creative contexts lead the two speakers to have a spontaneous discussion (see also more details in Section II.A). As a result, the recorded signals include speech and silence sections, as well as many different non-speech audible events such as those mentioned above. It also contains many face expressions and movements with or without sound production (see the related work of Macho *et al.* (2005)). Using this corpus, and starting from a very simple hypothesis –the lips of a given speaker should move when he/she is talking (or producing non-speech sounds), whereas they should not move (or move less) when he/she does not utter sounds– the distributions of static and dynamic lip parameters are provided for the two conditions. Those distributions show how dynamic lip parameters can be associated with non-silence sections (*i.e.* speech + non-speech audible events) *vs.* silence sections. Actually, the correspondence is not straightforward. Indeed, lip movements can occur during silence and conversely speech or non-speech oral production can occur with still lips. However, it is shown that a single dynamic lip parameter is more appropriate than static parameters for this characterization, and that temporal integration of the dynamic parameter values can improve the “separability” of non-silence sections *vs.* silence sections from lip information.

C. Application to automatic voice activity detection

Finally, a technological application of the study is considered: the possibility of using visual information to automatically detect sound production and silence sections in a given audio channel. Such an algorithm is called a Voice Activity Detector (VAD), and it is generally derived from audio information only. Among other applications, it can be used to drastically improve the performance of speech enhancement / separation techniques: *silence detection*, *i.e.* the detection of regions where the speaker of interest does not produce any sound, is used to identify properties of the noise or properties of the mixture configuration. These properties are then used to process the extraction of the speech signal of interest when it is detected as present in the mixture[1] (see, *e.g.*, Ephraïm and Malah (1984), Abrard and Deville (2003)). Various types of audio VAD have been studied, and they can achieve good performance even with a low signal-to-noise ratio (SNR) (Le Bouquin-Jeannès and Faucon, 1995; Sohn, *et al.*, 1999; Tanyer and Ozer, 2000; Ramírez *et al*., 2005). However these techniques are based on the analysis of the acoustic signal, and consequently their performance depends strongly on the environment noise. Generally the noise has to be considered as stationary or weakly non-stationary, and/or with a given power spectral density function (psd) or probability density function (pdf). Thus, when the noise is highly non-stationary with a low SNR (a concurrent speaker for example), the audio VAD performance considerably decreases. In this case, visual information could be very useful since it is completely independent of the acoustic environment[2]. For instance, in a previous study, De Cueto *et al.* (2000) used a basic Visual Voice Activity Detector (V-VAD) for detecting a speaker's speech activity in front of a computer. For this, either specific lip parameters or the average luminance of the mouth picture can be used (Iyengar and Neti, 2001). However, those studies are limited to the speaker's "intent-to-speak", useful for, e.g., turn-taking detection. The methods do not provide accurate segmentation of the content of a given speaker's sequences. More recently, Liu and Wang (2004) proposed a visual VAD based on Gaussian models. One Gaussian kernel was used to model the silence/non-speech sections and two kernels were used to model the speech sections[3]. However, little information is reported on the video processing, on the nature of the corpus that is used for setting and testing the V-VAD, and even on the visual information itself: it is not clear whether static or dynamic information is used. Also, the size of the experimental data is not compatible with real-life applications. The V-VAD proposed in the present paper specifically addresses these last remarks: it is based on "real-life" audiovisual data (and it is tested using these data), while remaining simple (given that lip shape parameters are available). Its efficiency is demonstrated by a series of detection scores (Receiver Operating Characteristics, ROC). As mentioned before, this V-VAD can be used in a speech enhancement system or a source separation system (see for instance Rivet *et al.*, 2007b, for a first application of V-VAD to the speech source separation problem).

This paper is organized as follows. Section II presents the method, beginning with a description of the audio-visual corpus (Section II.A) including the recording conditions and the definition of the video (lip) parameters used in this study; This is followed by the description of the audio (Section II.B) and video (Section II.C) processing applied to the data. The lip dynamic parameter used for silence vs. non-silence characterization and VAD is described in details in Section II.D. Section III presents the results of the study: in Section III.A, the audio content of the corpus in terms of silence vs. non-silence sections is presented. Then, Section III.B provides an analysis of the properties of the static and dynamic lip parameters in silence vs. non-silence sections. The performance of the proposed V-VAD in terms of ROC curves is given in Section III.C. Section IV is a conclusion section.

II. METHOD

A. Description of the audiovisual corpus

To describe and characterize lip movements in relation with speech/sound production or non production requires the acquisition of appropriate audiovisual data. An original audio-visual corpus was thus recorded and processed, consisting of a series of spontaneous dialogs between two male French speakers (JLS and LG). To obtain a set of conversation situations as natural as possible, several tasks were suggested to the speakers. These tasks were, *e.g.*, different interactive games such as answering as fast as possible to a word association problem, finding the solution of riddles, or playing language games. In all these tasks, the interaction between speakers was totally spontaneous, thus including spontaneous turn taking, interruptions, hesitations, and possible cross-overlapping between speakers. This led each of them to alternate between natural silence sections and speech sections of various sizes and contents. The corpus also contains many different kinds of audible and non-audible non-speech events, such as those mentioned in the introduction.

The two speakers were placed and recorded in separate rooms. They both had a microphone and a micro-camera fixed on a light helmet. The camera focused on the lip region to optimize the capture of labial information. Moreover, the speakers could hear and see each other, using headphones and a monitor screen in front of them with real-time video feedback. This was necessary to ensure “naturalness and conviviality” during the conversation. Automatic time-code generators were used for post-processing synchronization of all audio/video signals. Finally, these experimental settings enabled the conditions of a real face-to-face conversation to be simulated while the recorded audio signals (and of course the video signals) were perfectly separate. Illustrations of the recording session are given in Fig. 1.

The visual information extracted from this corpus consists of the time trajectories of two basic geometric parameters characterizing the lip contour (see Section I.A), namely inner width l_w and inner height l_h (Fig. 2). These parameters were extracted using the ICP “face processing system” (Lallouache, 1990), which is based on blue make-up, image thresholding with the Chroma-Key system, and contour tracking algorithms. The parameters were extracted every 20 ms (the video sampling frequency is 50 Hz), synchronously with the acoustic signal, which is sampled at 44.1 kHz. Thus, in the following, a signal *frame* is defined as a 20 ms section of acoustic signal together with a pair of lip parameters (l_w, l_h). A spontaneous audio-visual speech corpus for two speakers with a total duration of 40 min was finally obtained, representing 120,000 vectors of audio-visual frames per speaker.

B. Audio analysis and silence / non-silence labeling

The first phase of the corpus analysis consisted in the labeling of the 20 ms-frames (corresponding to the video sampling) as “silence frames” or “non-silence frames” based on the analysis of the audio signal and the dichotomy defined in the introduction: Silence frames are defined as signal frames with no sound produced at all, and non-silence frames contain speech and/or non-speech acoustic events. It is important to note that these definitions are given here for each speaker independently (obviously, a silence frame for one speaker can be simultaneous with a non-silence frame for the other speaker, since the two tracks are recorded separately). Silence frames are mainly present between phrase boundaries that result from conversation turn-taking, and also in more or less long pauses within one speaker’s “continuous” talk due to, e.g., hesitations.

The labeling into silence frames *vs.* non-silence frames was made semi-automatically with the algorithm proposed by Ramirez *et al.* (2004) and a manual verification. This algorithm measures the long-term spectral divergence between speech and environment noise and formulates the decision rule by comparing the long-term spectral envelope to the average noise spectrum, thus yielding a high discriminating decision rule and minimizing the average number of decision errors. The decision threshold is adapted to the measured noise. In our case, the environment noise was generally very low, and the results of this labeling were almost perfect. A manual verification of the entire corpus was made and a very small number of errors were corrected. It can be noted that very short silences corresponding to the time periods preceding the release of unvoiced plosives are not considered as silence frames, even though they may happen to be slightly greater than 20 ms. This is because of the nature of the audio detection algorithm that considers longer signal sections. Conveniently, this is coherent with the definition and processing of the temporal integration step that we propose in Section II.D.

C. Video pre-processing

As mentioned before, the extracted visual information is the time trajectory of the geometric parameters l_w and l_h characterizing the lip contour. The measures provided by the face processing system, although very accurate, are slightly noisy. Since a dynamic video parameter is calculated from the derivatives of the temporal trajectories, computed by a difference operator which is very sensitive to noise, the lip parameter trajectories have to be filtered (smoothed). This is not a trivial task for such signals, since labial parameter trajectories are highly non-stationary signals: slow variations in time can be followed by drastic changes, for instance when lips are closing. Therefore, it is difficult to remove noise in regions with slow variations while respecting the abrupt variations provided by natural lip movements. In our study, a technique based on spline functions was used. A basic version of this technique has been successfully used in a previous study using audio-visual corpora (Girin, 2004) and this process is refined here as follows.

The basic principle of the spline smoothing consists in locally fitting (noisy) data $x(i)$ with a cubic spline $s(t_i)$ defined as piecewise polynomial functions, where each piece is described using a cubic polynomial. The fitting is based on the minimisation of the following criterion:

$$f = p\sum_{j=1}^{J} w(j)\left|x(j)-s(t_j)\right|^2 + (1-p)\int\left(\frac{\partial s}{\partial t}\right)^2 dt \qquad (1)$$

The first term is a weighted least-square error between data and the spline model (the weights are given by $w(i)$) and the second term stands for the smoothness of the resulting curve. Balancing these two constraints is made possible by setting the parameter p at an appropriate value between 0 and 1. For instance, $p = 0$ produces a least squares straight line fit to the data, $p = 1$ produces a cubic spline interpolate, and intermediate values provide a trade-off between close fit and smoothness.

In the proposed video processing system, the non-stationary property of the lip movements is taken into account by adaptively tuning the p parameter according to the signal dynamics.

Relatively large p values must be used in time sections with high natural variations of the lip parameters to closely track these variations. On the contrary, relatively small p values must be used in quasi-stationary regions to adequately remove the noise. Thus the lip parameter signals l_w and l_h are segmented in time sections depending on the value of their local (sliding) variance $C(t) = 1/N\sum_{n=-N/2}^{N/2} v(t+n)^2$ with $N = 6$ ($v(t)$ represents a visual parameter (l_w or l_h), and t denotes the time index of 20ms-frames).

Each section is then fitted with a cubic spline whose parameter p is determined as a function of this variance. More specifically, this automatic smoothing process for each visual parameter $v(t)$ is the following:

- Compute for each frame the local variance $C(t)$.
- Search sections of consecutive frames with a variance $C(t)$ lower than a fixed threshold C_{min} defining a quasi-stationary signal section. Then all other frames are considered as non-stationary. This provides alternations of quasi-stationary sections and non-stationary sections with variable lengths.
- For each section i compute the mean of $C(t)$ over the section:

$$\overline{C}_i = \frac{1}{T_i}\sum_{t=t_i}^{t_i+T_i-1} C(t) \qquad (2)$$

(T_i denotes the size of the section i and t_i denotes the index of the first frame of the section) and compute p_i so that:

$$p_i = \begin{cases} p_{\min} & \text{if } \log_{10}\overline{C}_i < \lambda_{\min} \\ \dfrac{p_{\max}-p_{\min}}{\lambda_{\max}-\lambda_{\min}}\log_{10}\overline{C}_i + \dfrac{p_{\min}\lambda_{\max}-p_{\max}\lambda_{\min}}{\lambda_{\max}-\lambda_{\min}} & \text{if } \lambda_{\min} \leqslant \log_{10}\overline{C}_i \leqslant \lambda_{\max} \\ p_{\max} & \text{if } \log_{10}\overline{C}_i > \lambda_{\max} \end{cases} \qquad (3)$$

where the thresholds are fixed as:

$$\lambda_{\min} = \log_{10}\left(\frac{std(C)}{50}\right) \quad p_{\min} = 0.0001$$
$$\lambda_{\max} = \log_{10}\left(5std(C)\right) \quad p_{\max} = 0.8$$

Finally, the weights $w(i)$ of (1) are assumed to be equal to 1 for all data. This process is applied on each parameter $l_w(t)$ and $l_h(t)$ to obtain the smoothed visual parameters $\tilde{l}_w(t)$ and $\tilde{l}_h(t)$ [4]. An illustration of the results obtained with this process is given in Section III.B.

D. A dynamic lip parameter for silence vs. non-silence characterization and automatic silence detection

In Section I.A, we have briefly discussed the importance of the lip *movements* (as opposed to static lip shapes) for characterizing audio-visual speech. In a preliminary work, lip movements have been shown to be good candidates to characterize the opposition between silence and non-silence activity (Sodoyer *et al.*, 2006), the lip-shape variations being generally smaller in silence sections. Therefore, following this previous work, we chose to describe the lip shape movements with one dynamic parameter, summing the absolute values of the two lip parameter derivatives (Sodoyer *et al.*, 2006):

$$\pi(t) = \left|\frac{\partial \tilde{l}_w(t)}{\partial t}\right| + \left|\frac{\partial \tilde{l}_h(t)}{\partial t}\right| \tag{4}$$

Large $\pi(t)$ values indicate significant lip movements and should index non-silence frames, while low values corresponding to small lip movements (or no movement at all) should index silence sections. Note that this dynamic parameter exploits the complementarity between the two lip parameters for many speech sequences (*cf*. Fig. 3). Indeed, the variations of $\tilde{l}_w(t)$ may characterize rounding movements during which lip height may not change much; and vice versa, the variations of $\tilde{l}_h(t)$ may characterize opening/closing movements during which lip width may not change much. For example, in Fig. 3, the variations of the width parameter are larger than the variations of the height parameter between 278.5s and 278.8s, and the contrary occurs between 278s and 278.2s.

However, the situation is not so simple. On the one hand instantaneous large $\pi(t)$ values can correspond to local short lip movements in silence sections (*e.g.*, smiles, grimacing, funny faces or changes of the lips “rest position”). This is likely to produce silence detection errors (silence classified as non-silence). On the other hand, local lip stability within speech gestures can lead to low local $\pi(t)$ values providing false alarms (speech classified as silence). To overcome these problems, $\pi(t)$ values are then summed over time. Therefore, the parameter $\rho(t)$ is defined from the filtering of $\pi(t)$ as:

$$\rho(t) = h(t) * \pi(t) \tag{5}$$

with $h(t)$ being the truncated version of a first-order low-pass filter defined by:

$$h(t) = \frac{1}{\tau}\sum_{t=0}^{T-1}\exp\left(\frac{-t}{\tau}\right) \qquad (6)$$

where τ is the time constant of the filter and T is the number of integrated frames. These two parameters must be adequately chosen so that the filter significantly decreases the influence of isolated and accidental high $\pi(t)$ values in silence sections. On the other hand, the filter should not blur small but significant movements in non-silence sections. In our study, for the sake of simplicity, the filter length is fixed to T = 100 samples (or 2s) and several representative values for τ are tested in Section III.B (the τ value has the role of a memory factor over the past $\pi(t)$ values: the smaller τ, the shorter the memory).

Finally, the video-based automatic acoustic silence detection is achieved for each frame by comparing $\rho(t)$ to a threshold ρ_{th} that remains to be determined. Therefore, the problem can be formalized by the following hypotheses:

- H_s: The audio frame belongs to a silence section,
- H_{ns}: The audio frame belongs to a non-silence section.

Then, the audio frame index will respect the following rule:

$$\rho(t) \underset{H_{ns}}{\overset{H_s}{\lesseqgtr}} \rho_{th} \qquad (7)$$

i.e., if $\rho(t) < \rho_{th}$ the frame t is considered as silence, else it is considered as non-silence. This test is what is here referred to as Visual Voice Activity Detection (V-VAD).

III. QUANTITATIVE ASSESSMENT

A. Audio analysis results

The audio processing described in Section II.B has been applied to the corpus for each speaker (JLS and LG). As mentioned before, each frame (about 120,000 20-ms frames per speaker) was automatically labeled as "silence" or "non-silence" before a systematic manual verification. To illustrate the diversity of the corpus, Fig. 4 shows several audio sequences for both speakers. These examples illustrate the need for a distinction between silence and non-silence rather than speech *vs.* non-speech. Some audio sections with a significant amount of energy (non-silence), *e.g.*, Fig. 4(a) between 41.8s and 42.3s, Fig. 4(d) between 74.7s and 74.9s, or Fig. 4(e), between 26.5s and 27.1s, are not speech but rather grunts or murmurs. Table 1 presents some quantitative results, derived from the analysis, which provide a characterization of the corpus. The number of frames labeled as silence *vs.* non-silence is quite close for speakers JLS and LG (51% and 58% of the total corpus respectively). If a “silence section” is defined as a section composed of contiguous silence frames, and if a “non-silence section” is defined as a section composed of contiguous non-silence frames, 691 silence sections and 695 non-silence sections are obtained for speaker JLS, and 603 silence sections and 607 non-silence sections are obtained for speaker LG, with respective average time lengths of 1.73s and 1.93s for the first speaker and, 2.55s and 1.85s for the second one. The corresponding standard deviations are quite high (the section length ranges from one to more than 2000 frames, that is 40s), illustrating the diversity of dialog situations. Fig. 5 shows the duration histograms of silence and non-silence sections. In both cases, more than 90 % of the sections have a duration lower than 4 s.

B. Video characterization of silence vs. non-silence

For each speaker, the labial parameters $l_w(t)$ and $l_h(t)$ were smoothed with the pre-processing described in section II.C. Fig. 6 shows the results of this process. It can be seen that the adaptive spline filter efficiently removes the measurement noise: slowly varying sections seem correctly smoothed, whereas fast parameter variations in highly non-stationary sections are preserved. Fig. 7 shows the distribution of the resulting lip parameters for both speakers, separately for the audio silence frames and the non-silence frames. First, differences between the distributions for the two speakers can be noticed. These differences are simply due to inter-individual differences in lip shapes and gestures. Despite these differences, the two distributions have similar shapes in the non-silence context (Fig. 7(a) and Fig. 7(c)). For each speaker, the resulting organization of the labial space is classical for speech configurations (Benoît *et al.* 1992; Robert-Ribes *et al.*, 1998), assuming that the additional non-speech gestures do not smear the global trends. For example, we can distinguish closed lip shapes ($\tilde{l}_w(t) = 0$ and $\tilde{l}_h(t) = 0$) corresponding to bilabials in any vocalic context, rounded lip shapes (*e.g.*, [y], [u], at around $\tilde{l}_w(t) = 2$ cm and $\tilde{l}_h(t) = 0.25$ cm, and consonants in rounded contexts), spread lip shapes (*e.g.*, [i], at around $\tilde{l}_h(t) = 3.5$ cm and $\tilde{l}_h(t) = 0.6$ cm, and consonants in spread contexts) and open lip shapes (*e.g.*, [a], at around $\tilde{l}_w(t) = 3.5$ cm and

$\tilde{l}_h(t)$ = 1 cm, see also Fig. 3, and consonants in open contexts). Notice that closed lip shapes represent 10% of non-silence frames for both speakers (see Table 1). This is a typical example of the difficulty to associate a given lip shape to a given audio class: in this specific case, a speaker actually spoke or emitted sounds with his mouth shut (during short periods). Now, let us consider the visual parameter distribution associated with audio silence, in Fig. 7(b) and Fig. 7(d). These figures show that an important subset of visual parameters corresponding to silence frames is located in a sub-region within the general set of speech shapes displayed in Fig. 7(a) and Fig. 7(c). Besides, another important subset of lip configurations is grouped around the origin, which corresponds to closed lips. Table 1 however shows that closed lip shapes represent only 27% to 30% of the lip shapes associated to silence frames. This is much more than the 10% proportion in non-silence frames, but quite far from the totality of silence frames. Altogether, it appears that closed lip shapes are present in both distributions and thus cannot be systematically associated with a silence frame. More generally, since most values of the distribution of static visual parameters ($\tilde{l}_w(t)$, $\tilde{l}_h(t)$) associated to either silence frames or non-silence frames are located in the same region, this information is not sufficient to characterize audio silence *vs.* non-silence. This confirms the need for a dynamic characterization of lip gestures.

A first illustration of this is given in Fig. 8, which provides the same plots as Fig. 7, but for the derivatives of the parameters (on a log scale for a better concentration of the values). We can see that, although still overlapping, the silence and non-silence distributions are globally much better separated than previously, with the distributions for non-silence frames being concentrated in higher parameter values than for silence frames. Also, the differences in the distributions between the two speakers seem to be much smaller in this case than in the static case, for both silence and non-silence frames.

Fig. 9 displays the distribution (here as an histogram) of the dynamic parameter $\rho(t)$ for the entire corpus respectively for speaker JLS (left column) and speaker LG (right column), and for four values of the time constant τ corresponding to the summation of 1 frame (that is no actual temporal summation), 5 frames (100ms integration), 10 frames (200ms) and 100 frames (2s). The underlying goal is to tune the temporal-integration window so that the distributions of $\rho(t)$ corresponding to the silence sections (the histogram plotted in black in Fig. 9) and to the non-silence sections (the histogram plotted in white) are as separate as possible. Each of these two distributions is grossly distributed among two classes: the first one is a peak on the left part of the figure corresponding to no lip movement (including of course stable closed lips), and the second one is a kernel on the right part of the figure corresponding to the presence of lip movements. The two kernels associated with silence frames (plotted in black) and non-silence frames (plotted in white) are centered on different locations, the non-silence kernel being to the right of the silence kernel. This confirms that non-silence sections are generally associated with larger/faster movements of the lips than silence sections. However, the two kernels are strongly overlapping for the 1-frame integration, as shown by Fig. 9(a) and Fig. 9(e), since short lip movements can occur during audio silences. Furthermore, the distribution peak associated to stable closed lips on the left part of these figures contains a large contribution of non-silence frames, since short stable lip shapes can occur during speech/sound activity. An optimal temporal integration window is required, which should provide the best separation of these kernels, while reducing the proportion of no-movement values associated with non-silence frames. Too large a time constant (as in Fig. 9(d) and Fig. 9(h)), while successfully addressing this last point, mixes the silence and non-silence kernels too much, losing the discrimination between silence and non-silence audio frames for moving lips. However, the histograms plotted in Fig. 9(c, g) show that a suitable time summation around 5-10 frames (100-200 ms) can largely improve the discrimination between silence and non-silence sections (actually the optimal value is likely to be closer to 5 than to 10): in this case, the white portion of the peak at the origin is quite small and the black and white kernels are better separated than in the other configurations. Notice finally that the dynamic parameter $\rho(t)$ provides less difference between speakers than the static labial parameters, as was already observed in Fig. 8. This could be important for a future multi-speaker application.

C. Automatic video-based silence detection

The proposed V-VAD of Section II.D was tested on the 120,000 frames of the corpus, and for the different settings of the time integrations: 1 frame (instantaneous case), 5, 10, 20 and 100 frames. In each case, the results of automatic silence frame detection using the V-VAD were compared with the reference labels provided by the acoustic semi-automatic identification process presented in Section II.B. This test has been done for each speaker.

Fig. 10 shows an example of silence detection. This figure represents the time trajectory of the lip parameters $\tilde{l}_w(t)$ and $\tilde{l}_h(t)$ (Fig. 10(a) and Fig. 10(b)), of their respective derivatives (Fig. 10(c) and Fig. 10(d)), and of the dynamic parameters $\pi(t)$ and $\rho(t)$ with their corresponding detection thresholds (Fig. 10(e) and Fig. 10(f)), for about 7 s of signal produced by speaker JLS. Fig. 10(g) represents the corresponding speech waveform with the detected and reference silence regions. This figure illustrates the different possible relations between visual and acoustic data: movement of the lips in non-silence (*e.g.*, from 29.7s to 30.6s) and in silence (*e.g.*, just before 31.5s, or between 32s and 32.3s), non-movement of the lips in silence with opened lips (*e.g.*, from 31.2s to 31.4s) and closed lips (from 31.5s to 31.9s), and non-movement in non-silence (from 30.9s to 31.1s). The V-VAD, adequately tuned ($\tau = 20$), performs quite well. The silence section of this sequence has been detected. Obviously, the V-VAD fails to avoid a false detection between 31s and 31.2s, but this is a tough configuration: part of this mistakenly detected section is a long non-silence section with still lip shape, corresponding to a drawling sentence ending. Moreover, the V-VAD has shrunk the actual silence section. But on the other hand, it discards several possible false detections in the speech section between 32.5s and 36s, in spite of both closed lips sections and small movements in some regions.

More general results are presented in Fig. 11 as Receiver Operating Characteristics (ROC). These curves represent the percentage of correct silence detection (defined as the ratio between the number of detected silence frames and the actual number of silence frames) as a function of the percentage of false silence detection (defined as the ratio between the number of non-silence frames detected as silence frames and the actual number of non-silence frames). To obtain those curves, the threshold ρ_{th} was varied between the minimum and the maximum of $\rho(t)$ (however, when using the V-VAD, one would set ρ_{th} to a fixed value ensuring a good trade-off between hit rate and false alarm, possibly using the ROC curves as charts). It can be seen from those curves that the benefit of low-pass filtering the parameter $\rho(t)$ is significant. By decreasing the influence of short stable periods in actual speech or sound production, it enables the false silence detection ratio to be decreased significantly. Symmetrically, by decreasing the influence of short/small lip movements in silence, it improves the silence detection ratio. The time integration must be set carefully. When no time integration is performed, the false silence detection scores are moderate (*e.g.*, the point 20%-80% for speaker JLS, and 22%-80% for speaker LG). On the contrary, too large a time integration ($\tau = 100$ frames corresponding to 2s) dramatically decreases the silence detection ratio. Finally, the ROC performances are significantly improved with suitable time integration. For instance, using $\tau = 5$ frames (corresponding to 100 ms) efficiently decreases the false silence detection ratio without decreasing the silence detection ratio: ROC scores of 12%-80% and 15%-80% are obtained for speaker JLS and speaker LG respectively.

As a complementary result, Fig. 12 shows the ROC curves obtained when $l_w(t)$ and $l_h(t)$ are used in (4), *i.e.*, unfiltered visual parameters, instead of $\tilde{l}_w(t)$ and $\tilde{l}_h(t)$, to compute $\rho(t)$ with (5). In this case, lower performances are obtained, which confirms the importance of the pre-processing. Moreover, the role of integration is more important in this case because it also reduces the influence of the measurement noise coming from the lip parameters extraction system. This explains that the difference between the results of Fig. 11 and Fig. 12 is particularly important if no integration is performed (*e.g.*, 37%-80% in the no-integration case compared to 17%-80% with adequate integration). The results with temporal integration are quite close with or without pre-processing for speaker JLS, although they are better with the pre-processing than without the pre-processing for speaker LG. This seems to be due to greater measurement noise for this last speaker.

IV. CONCLUSION

This paper had two objectives. The first one was to describe the recording and processing of an audiovisual corpus in natural interaction situations. The second objective was to use this corpus to characterize the visual information provided by a speaker’s lips during the different dialog phases, with a particular focus on silence sections. An automatic simple and efficient Visual Voice Activity Detector was derived from this analysis.

Regarding the first objective, let us recall that the corpus contains about 40 min of signal, providing a rich set of audiovisual data for two speakers in a realistic situation of spontaneous dialog (in French). This corpus is dedicated to fundamental studies in speech and language sciences, as well as to the assessment of audio-visual speech processing systems. The design of such a corpus is not a straightforward task. It requires specific recording equipment and protocol. In addition, as was pointed out in this paper, the pre-processing of the video data is not trivial (although it can be easily implemented after adequate settings). This corpus can be downloaded free of charge from http://www.icp.inpg.fr, assuming it is used for scientific / non-profit purposes.

Regarding the second objective, the results show that the instantaneous lip shapes in silence and non-silence frames are largely overlapping. Consequently, such straightforward information cannot be efficiently used for silence *vs.* non-silence automatic classification of speech sequences. In contrast, lip movements can provide adequate information: A single dynamical parameter processed with suitable temporal integration and threshold has been shown to be appropriate for efficient silence (*vs.* non silence) detection. The detection scores have shown that the resulting Visual VAD (actually a visual *silence* detector) can be exploitable in real speech processing applications like enhancement, source separation or recognition in noise, with, *e.g.*, a 12% false alarm rate *vs.* an 80% hit rate. It is of primary importance to remember that these performance scores are completely independent of the acoustic environment, a property that is not ensured by classical acoustic VAD. Note finally that, in the perspective of a "real world" implementation, the blue make-up used for labial information extraction is not a limitation of the proposed method. In a recent study (Aubrey *et al.*, 2007), it has been shown that the dynamic information provided by (6) is equivalent (in terms of detection scores) to the information provided by a retina model applied on raw black and white images of the lip region, with natural lips (*i.e.*, without make-up).

Further investigations will be conducted to increase the V-VAD performance. They could incorporate an adaptive decision threshold taking into account the image quality and/or the inter-speaker variability. Another perspective is to use both video and audio information together to increase the detection performance, either taking a decision from a fusion of the decisions provided independently by audio and video information, or using both sources of information to feed a single decision process. This would lead to the design of an Audio-Visual VAD, which seems to us an important outcome for future developments in audiovisual speech processing systems. The visual VAD that has been presented in this study provides a good basis for such development.

ENDNOTES

1. Note that this explains why all throughout the paper we consider the distinction between *silence sections* and *non-silence sections* (including speech and non-speech audible events), rather than the distinction between speech and non-speech (including silence and non-speech audible events). Accordingly, the term *Voice Activity* is to be understood as covering speech and non-speech audible events (while voice *inactivity* would correspond to silence). The term VAD is a usual denomination in the speech processing literature.

2. Yet a dependence can be found by considering "the Lombard effect" (Lombard, 1911; Lane and Tranel, 1971): The speaker may increase his/her articulatory efforts (and thus modify the speech characteristics) to improve communication efficiency in noise. This does not reduce the interest of the visual speech information (on the contrary, the movements of the visible articulators may be exaggerated by the Lombard effect).

3. The authors prefer to classify between speech and non-speech sections rather than between silence and non-silence sections as we do, even if it seems less appropriate for use in enhancement/separation applications.

4. Actually, it is not applied in regions where the parameters are equal to zero, or more specifically, the zero value in those regions is not modified, since (i) the zero signal is not noisy, and (ii) this avoids unwanted oscillations or overshoots of the spline-filtered parameters after fast lip closing or before fast lip opening regions. In practice, implementing this precaution is a trivial task.

REFERENCES

Abrard, F., and Deville, Y. (**2003**). “Blind separation of dependent sources using the “time-frequency ratio of mixture” approach,” in Proc. International Symposium on Signal Processing and its Applications (ISSPA), Paris, France, 81–84.

Abry, C., and Boë, L.J. (**1986**). “Laws for lips,”. Speech Communication, **5**, 97-104.

Aubrey, A., Rivet, B., Hicks, Y., Girin, L., Chambers, J., and Jutten, C. (**2007**). “Comparison of appearance models and retinal filtering for visual voice activity detection,” in Proc. European Signal Processing Conference (EUSIPCO), Poznan, Poland.

Bailly, G., and Badin, P. (**2002**). “Seeing tongue movements from outside,” in Proc. International Conference Spoken Language Processing (ICSLP), Denver, Colorado, 1913–1916.

Bailly, G., Berard, M., Elisei, F., and Odisio, M. (**2003**). “Audiovisual speech synthesis,” International Journal of Speech Technology, **6**(4), 331–346.

Barker, J.P., and Berthommier, F. (**1999**). “Estimation of speech acoustics from visual speech features: a comparision of linear and non-linear models,” in Proc. Audio-Visual Speech Processing (AVSP), Santa Cruz, California, 112–117.

Benoit C., Lallouache T., Mohamadi T., and Abry C. (**1992**). “A Set of French Visemes for Visual Speech Synthesis”, in *Talking machines:Theories, Models, and Designs* edited by Bailly G., Benoit C., and Sawallis T.R. (North-Holland, Amsterdam), 485–504.

Benoît, C., Mohamadi, T. and Kandel, S. (**1994**). “Effects of phonetic context on audio-visual intelligibility of French,” J. Speech and Hearing Research, **37**, 1195–1293.

Benoît, C., Guiard-Marigny, T., Le Goff, B., and Adjoudani, A. (**1996**). “Which components of the face humans and machines best speechread ?” in *Speechreading by man and machine: Models, Systems and Applications* edited by Stork D.G. and Hennecke M.E., (Springer, NATO ASI Series), 315–328.

Bertelson, P. (**1999**). “Ventriloquism: a case of crossmodal perceptual grouping,” in *Cognitive Contributions to the Perception of Spatial and Temporal Events*, edited by G. Aschersleben, T. Bachmann, and J. Müsseler (Elseviers, Amsterdam), 347–362.

Calvert, G.A. & Campbell, R. (**2003**). “Reading speech from still and moving faces : the neural substrates of visible speech," J. of Cognitive Neuroscience, **15**, 57-70.

Campbell, N. (**2007**). "Approaches to conversational speech rhythm: speech activity in two-person telephone dialogues," in Proc. International Congres of Phonetic Sciences (ICPhS), Sarrebrücken, Germany, 343-348.

Cosi, P., Fusaro, A., and Tisato, G. (**2003**). "LUCIA: a new Italian talking-head based on a modified Cohen-Massaro's labial coarticulation model," in Proc. European Conference on Speech Communication and Technology (EuroSpeech), Geneva, Switzerland, 2269–2272.

De Cueto, P., Neti, C. and Senior. A. W. (**2000**). "Audio-visual intent-to-speak detection in human-computer interaction," in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Istanbul, Turkey, 2373–2376.

Deligne, S., Potamianos, G., and Neti, C. (**2002**). "Audio-visual speech enhancement with AVCDCN (AudioVisual Codebook Dependent Cepstral Normalization)," in Proc. International Conference Spoken Language Proc. (ICSLP), Denver, Colorado, USA, 1449–1452.

Ephraim, Y., and Malah, D. (**1984**). "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Trans. Acoust., Speech, Signal Processing, **32**, 1109–1121.

Erber, N. P. (**1975**). "Auditory-visual perception of speech," J. Speech and Hearing Disorders, **40**, 481–492.

Gibert, G., Bailly, G., Beautemps, D., Elisei, F., and Brun, R. (**2005**). "Analysis and synthesis of three-dimensional movements of the head, face, and of a speaker using cued speech," J. Acoust. Soc. Am., **118**, 1144–1153.

Girin, L., Schwartz, J.-L., and Feng, G. (**2001**). "Audio-visual enhancement of speech noise," J. Acoust. Soc. Am., **109**, 3007–3020.

Girin, L. (**2004**). "Joint matrix quantization of face parameters and LPC coefficients for low bit rate audiovisual speech coding," IEEE Trans. Speech and Audio Processing, **12**, 265–276.

Goecke, R., and Millar, J. B. (**2003**). "Statistical analysis of relationship between audio and video speech parameters australian English," in Proc. Audio-Visual Speech Processing. (AVSP), Saint-Jorioz, France, 133–138.

Grant, K. W., and Seitz P. (**2000**). "The use of visible speech cues for improving auditory detection of spoken sentences," J. Acoust. Soc. Am., **108**, 1197–1208.

Huang, J., Liu, Z., Wang, Y., Chen, Y., and Wong, E. (**1999**). "Integration of multimodal feature for video scene classification based on HMM," in Proc. Workshop Meeting on Multimedia Signal Processing (MMSP), Copenhagen, Denmark, 53–58.

Iyengar, G., and Neti. C. (**2001**). "A vision-based microphone switch for speech intent detection," in Proc. Workshop at International Conference on Computer Vision (ICCV) on Recognition, Analysis and Tracking of Face and Gestures in Real Time Systems (RATFG-RTS), Vancouver, Canada, 101–105.

Jiang, J., Alwan, A., Keating, P. A., Auer, E. T., and Bernstein, L. E. (**2002**). "On the relationship between face movements, tongue movements and speech acoustics," Eurasip Journal on Applied Signal Processing, **11**, 1174–1188.

Kim, J., and Davis, C. (**2004**). "Investigating the audio-visual speech detection advantage," Speech Communication, **44**(1–4), 19–30.

Lallouache, T. (**1990**). "Un poste visage-parole. Acquisition et traitement des contours labiaux (*a device for the capture and processing of lip contours*)," in Proc. XVIII Journées d'Étude sur la Parole (JEP), Montréal, Canada, 282–286 (in French).

Lane, H., and Tranel, B. (**1971**). "The Lombard sign and the role of hearing in speech," J. Speech and Hearing Research, **14,** 677–709.

Le Bouquin-Jeannès, R., and Faucon, G. (**1995**). "Study of a voice activity detector and its influence on a noise reduction system," Speech Communication, **16**, 245–254.

Liu, P., and Wang, Z. (**2004**). "Voice activity detection using visual information," in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Montreal, Canada, 609–612.

Lombard, E. (**1911**). "Le signe de l'élévation de la voix (*the sign of voice rise*)," Annales des maladies de l'oreille et du larynx **37**, 101–119, (in French).

Macho, D., Padrell, J., Abad, A., Nadeu, C., Hernando, J., McDonough, Wölfel, M., Klee, U., Omologo, M., Brutti, A., Svaizer, P., Potamianos, G., and Chu, S.M. (**2005**). "Automatic speech activity detection, source localization and speech recognition on the CHIL seminar corpus," in International Conference on Multimedia and Expo (ICME), Amsterdam, The Netherlands, 876–879.

McGurk, H., and McDonald J. (**1976**). “Hearing lips and seeing voices,” Nature, **264**, 746–748.

Munhall, K. G., Gribble, P., Sacco, L., and Ward, M. (**1996**). “Temporal constraints on the McGurk Effect.” Perception and Psychophysics, **58**, 351-362.

Munhall, K. G., Servos, P., Santi, A. & Goodale, M. (**2002**). “Dynamic visual speech perception in a patient with visual form agnosia.” Neuroreport, **13**(14), 1793-1796.

Munhall, K. G., and Vatikiotis-Bateson, E. (**1998**). “The moving face during speech communication.” in *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* edited by Campbell R., Dodd B. and Burnham D. (Psychology Press, London), 123–139.

Petajan, E. D. (**1984**). “Automatic lipreading to enhance speech recognition,” in Proc. Global Telecommunications Conference (GLOBCOM), Atlanta, Georgia, 265–272.

Potamianos, G., Neti, C., and Gravier, G. (**2003**). “Recent advances in the automatic recognition of visual speech,” in Proc. IEEE, **91**(9), 1306–1326.

Potamianos, G., Neti, C., and Deligne, S. (**2003b**). “Joint audio-visual speech processing for recognition and enhancement,” in Proc. Audio-Visual Speech Processing (AVSP), Saint-Jorioz, France, 95–104.

Ramirez, J., Segura, J. C., Benıtez, C., de la Torre, A., and Rubio, A. (2004). “Efficient voice activity detection algorithms using long-term speech information,” Speech Communication, **42**, 271–287.

Ramírez, J., Segura, J. C., Benítez, C., García, L., and Rubio, A. (**2005**). “Statistical voice activity detection using a multiple observation likelihood ratio test,” IEEE Signal Processing Letters, **12**(10), 689–692.

Rao, R., and Chen, T. (1996). “Cross-modal predictive coding for talking head sequences,” in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Atlanta, Georgia, 2058–2061.

Rivet, B., Girin, L., and Jutten. C. (**2007**). “Mixing audiovisual speech processing and blind source separation for the extraction of speech signals from convolutive mixtures,” IEEE Trans. on Speech and Audio Processing, **15**(1), 96–108.

Rivet, B., Girin, L., and Jutten. C. (**2007b**). “Visual voice activity detection as a help for speech source separation from convolutive mixtures,” Speech Communication, **49**(7/8), 667–

677.

Robert-Ribes, J., Schwartz, J. L., Lallouache, T., and Escudier, P. (**1998**). “Complementary and synergy in bimodal speech: auditory, visual, and audio-visual identification of French oral vowels in noise,” J. Acoust. Soc. Am., **6**, 3677–3689.

Rosenblum, L. D., and Saldana, H. M. (**1996**). “An audiovisual test of kinematic primitives for visual speech perception.” J. Experimental Psychology: Human Perception and Performance, **22**(2), 318–331.

Rosenblum, L. D., Johnson, J. A., and Saldana, H. M. (**1996**). “Visual kinematic information for embellishing speech in noise.” J. Speech and Hearing Research, **39**(6), 1159-1170.

Schwartz, J. L., Berthommier, F., and Savariaux, C. (**2004**). “Seeing to hear better: Evidence for early audio-visual interactions in speech identification,” Cognition, **93**, 69–78.

Sodoyer, D., Girin, L., Jutten, C., and Schwartz, J. L. (**2002**). “Separation of audio-visual speech sources: A new approach exploiting the audiovisual coherence of speech stimuli,” Eurasip Journal on Applied Signal Processing, **11**, 1165–1173.

Sodoyer, D., Girin, L., Jutten, C., and Schwartz, J. L. (**2004**). “Further experiments on audio-visual speech source separation,” Speech Communication, **44**(1–4), 113–125.

Sodoyer, D., Rivet, B., Girin, L., Jutten, C., and Schwartz, J. L. (**2006**). “An analysis of visual speech information applied to voice activity detection”, in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toulouse, France, 601–604.

Sohn, J., Kim, N. S., and Sung, W. (**1999**). “A statistical model based voice activity detection,” IEEE Signal Processing Letters, **6**(1), 1–3.

Sumby, W. H., and Pollack I. (**1954**). “Visual contribution to speech intelligibility in noise,” J. Acoust. Soc. Am., **26**, 212–215.

Summerfield, Q. (**1987**). “Some preliminaries to a comprehensive account of audio-visual speech perception.” in *Hearing by eye: The psychology of lip-reading* edited by Dodd B. and Campbell R. (Erlbaum, London), 3–51.

Summerfield, Q. (**1979**). “Use of visual information for phonetic perception,” Phonetica, **36**, 314–331.

Tanyer, S. G., and Ozer, H. (**2000**). “Voice activity detection in nonstationary noise,” IEEE Trans. on Speech and Audio Processing, **8**(4), 478–482.

Thomas, S. M., and Jordan, T. R. (**2004**). “Contributions of oral and extraoral facial movement to visual and audiovisual speech perception,” J. Experimental Psychology, **30**(5), 873–888.

Wang, W., Cosker, D., Hicks, Y., Sanei, S., and Chambers, J. A. (**2005**). “Video assisted speech source separation,” in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Philadelphia, USA, 425–428.

Yehia, H., Rubin, P., and Vatikiotis-Bateson, E., (**1998**). “Quantitative association of vocal-tract and facial behavior,” Speech Communication, **26**, 23–43.

Yehia, H., Kuratate, T., and Vatikiotis-Bateson, E. (**2000**). “Facial animation and head motion driven by speech acoustics”, in Proc. Seminar on Speech Production: Models and Data & CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling, Kloster Seeon, Germany, 265–268.

TABLE

		JLS	LG
Number of **silence sections**		695	603
	Mean duration	1,73s	2,55s
	Standard deviation of duration	2,13s	3,49s
	Minimum duration	0,02s	0,04s
	Maximum duration	22,98s	41,98s
Number of **non-silence sections**		691	607
	Mean duration	1,93s	1,85s
	Standard deviation of duration	2,08s	1,85s
	Minimum duration	0,02s	0,02s
	Maximum duration	16,7s	12,8s
N	Total number of frames	119996	119996
N_s	Number of silence frames	61373 (51% of N)	69162 (58% of N)
N_{ns}	Number of non-silence frames	58623 (49% of N)	50834 (42% of N)
N_z	Number of frames with $\tilde{l}_w(t)$ and $\tilde{l}_h(t)$ null	22658 (19% of N)	26249 (22% of N)
N_{zns}	Number of non-silence frames with $\tilde{l}_w(t)$ and $\tilde{l}_h(t)$ null	5915 (10% of N_{ns})	4908 (10% of N_{ns})
N_{zs}	Number of silence frames with $\tilde{l}_w(t)$ and $\tilde{l}_h(t)$ null	16743 (27% of N_s)	21341 (31% of N_s)

TABLE I. Characteristics of the audio-visual corpus processed in this study. The frame size is 20 ms. The data in this table are derived from the semi-automatic audio process of Ramirez *et al.* (2004), with manual verification.

FIGURE CAPTIONS

FIG. 1. Illustrations of the audio-visual corpus recording session. The two speakers are in separate rooms. A specially-designed equipment is used for the real-time transmission of audio and video signals between the speakers, as well as the recording of these signals.

FIG. 2. The lip parameters used in this study: inner lip height (l_h) and inner lip width (l_w).

FIG. 3. Example of lip parameter trajectories: (top) inner width parameter, (middle) inner height parameter, (bottom) corresponding acoustic signal.

FIG. 4. Examples of sounds present in the spontaneous speech corpus. (a) and (b): typical hesitation sound in French (“euh”, a long [ø]; included in the sequence in (b)); (c): sound of “Mmmm…”; (d): snap of the lips before speech; (e): respiration intake; (f): laugh.

FIG. 5. Histograms of the time length (in seconds) of (top) silence sections, and (bottom) non-silence sections, for speaker LG.

FIG. 6. A lip width parameter trajectory filtered with the adaptive spline technique. Top: raw parameter; bottom: smoothed parameter. The slowly varying sections are efficiently smoothed while the abrupt changes are preserved.

FIG. 7. Distribution of the visual parameters for the two speakers JLS (top: (a) and (b)) and LG (bottom: (c) and (d)) and for the non-silence frames (left: (a) and (c)) and silence frames (right: (b) and (d)).

FIG. 8. Distribution of the (absolute values of the) derivatives of the lip parameters (on a log scale: $\tilde{\delta}_h = \log_{10}\left|\frac{\partial \tilde{l}_h}{\partial t}\right|$ and $\tilde{\delta}_w = \log_{10}\left|\frac{\partial \tilde{l}_w}{\partial t}\right|$) for the two speakers JLS (top: (a) and (b)) and LG

(bottom: (c) and (d)) and for the non-silence frames (left: (a) and (c)) and silence frames (right: (b) and (d)).

FIG. 9. Distribution of $\log_{10}(\rho(t))$ for the two speakers JLS (left column) and LG (right column) and for different configurations of the time integration. Note that the value $\rho(t) = 0$ (no movement) has been arbitrarily fixed to 10^{-4} for visualization of the origin.

FIG. 10. Silence detection on a sequence of the recorded corpus. (a) and (b): Static lip parameters $\tilde{l}_w(t)$ and $\tilde{l}_h(t)$; (c) and (d): Their derivatives (absolute values); (e) and (f): Instantaneous detection parameter $\pi(t)$ and integrated detection parameter $\rho(t)$ (for $\tau = 20$ frames = 400 ms), on a log-scale; the dotted and dashed lines are respectively the threshold for $\pi(t)$ and for $\rho(t)$; (g): Acoustic signal with silence reference (solid line), frames detected as silence using $\pi(t)$ (dotted line), and frames detected as silence using $\rho(t)$ (dashed line).

FIG. 11. ROC silence detection curves for the two speakers JLS (left) and LG (right). For each speaker, five integration durations of the visual parameter $\rho(t)$ are used: No integration (dotted line), 100 ms ($\tau = 5$, solid line), 200 ms ($\tau = 10$, dash-dot line), 400 ms ($\tau = 20$, dashed line) and 2 s ($\tau = 100$, small dashed line).

FIG. 12. ROC silence detection curves for the two speakers JLS (left) and LG (right). Here, the visual parameter $\rho(t)$ has been computed (using (5)) with unfiltered lips parameters l_h and l_w in (4). For each speaker, five integration durations of the visual parameter $\rho(t)$ are used: No integration (dotted line), 100 ms ($\tau = 5$, solid line), 200 ms ($\tau = 10$, dash-dot line), 400 ms ($\tau = 20$, dashed line) and 2 s ($\tau = 100$, small dashed line).